Valid confidence intervals post-model-selection

François Bachoc

Institut de Mathématiques de Toulouse
Université Paul Sabatier

Joint work with **Hannes Leeb** (Vienna), **Benedikt Pötscher** (Vienna), **David Preinerstorfer** (St. Gallen) and **Lukas Steinberger** (Vienna)

Rencontres Statistiques Lyonnaises
October 2023

Outline

Data generating process

Data

$$Y = \mu + U.$$

- Y of size $n \times 1$: observation vector.
- μ of size $n \times 1$: unknown mean vector.
- $U \sim \mathcal{N}(0, \sigma^2 I_n)$.
- σ^2 known in the first three sections for simplicity of exposition.

The linear model

- Design matrix X of size $n \times p$.
  - $p < n$ and X is full (column)-rank in slides 4-7.
- A column $=$ an explanatory variable.
- Let $\mathrm{span}(X)$ be the linear subspace of $\mathbb{R}^n$ generated by the columns of X.
- Projection of observation vector Y on $\mathrm{span}(X)$:

$$P_X(Y) = X(X'X)^{-1}X'Y.$$

- Called least square estimation because

$$P_X(Y) = \underset{v \in \mathrm{span}(X)}{\mathrm{argmin}} \|Y - v\|^2.$$

- $P_X(Y) = X\hat{\beta} =$ linear combinations of columns of X with coefficients given by

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Distributional properties of the linear model

- $\hat{\beta}$ is a Gaussian vector $\longrightarrow$ linear combination of Y.
- Expectation:

$$\begin{aligned}\mathbb{E}(\hat{\beta}) =&(X'X)^{-1}X'\mathbb{E}(Y)\\ =&(X'X)^{-1}X'\mu,\end{aligned}$$

so

$$\begin{aligned}\mathbb{E}(X\hat{\beta}) =&X(X'X)^{-1}X'\mu\\ =&P_X(\mu).\end{aligned}$$

- Covariance:

$$\begin{aligned}\mathrm{cov}\left(\hat{\beta}\right) =&(X'X)^{-1}X'\mathrm{cov}(Y)X(X'X)^{-1}\\ =&(X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1}\\ =&\sigma^2(X'X)^{-1}.\end{aligned}$$

Interpretations for the linear model: well-specified case

Well-specified setting

There exists a $p \times 1$ vector β_0 such that

$$\mu = \mathbb{E}(Y) = X\beta_0.$$

Then β_0 is the target and for $j \in \{1, \dots, p\}$,

- $(\beta_0)_j = 0 \implies$ variable j has no effect on the response,
- $(\beta_0)_j > 0 \implies$ variable j has a positive effect on the response,
- $(\beta_0)_j < 0 \implies$ variable j has a negative effect on the response.

Here effect $\approx$ causality.
Note that $\mathbb{E}(\hat{\beta}) = \beta_0 \implies \hat{\beta}$ is unbiased.

Interpretations for the linear model: misspecified case

Misspecified setting

Now, $\mu \notin \mathrm{span}(X)$.
But we can define

$$P_X(\mu) = X(X'X)^{-1}X'\mu = X\beta^\star.$$

Then $\beta^\star$ is the target and for $j \in \{1, \ldots, p\}$,

- $(\beta^\star)_j = 0 \Longrightarrow$ variable j has no effect on the response,
- $(\beta^\star)_j > 0 \Longrightarrow$ variable j has a positive effect on the response,
- $(\beta^\star)_j < 0 \Longrightarrow$ variable j has a negative effect on the response.

Here effect $\approx$ dependence / predictive power.
Note that $\beta^* = \mathbb{E}(\hat{\beta})$.

Linear models with variable selection

- Design matrix X of size $n \times p$.
  - $p < n$ or $p \geq n$.
- Universe $\mathcal{M}$ of models/submodels.

$$\mathcal{M} \subseteq \{M \subseteq \{1, \ldots, p\}\}.$$

  - Each $M \in \mathcal{M}$ is a set of selected columns of X.
  - Write $|M|$ for the cardinality of M.
  - Write $X[M]$ of size $n \times |M|$: only the columns of X that are in M.
- Restricted least square estimator

$$\hat{\beta}_M = \left(X'[M]X[M]\right)^{-1} X'[M]Y.$$

  - For $M \in \mathcal{M}$.
  - Assuming $X[M]$ has full column rank for $M \in \mathcal{M}$.
  - Implies $|M| \leq n$.

$\Longrightarrow$ We consider subsets of selected variables and construct linear models from them.

Examples of universes of models $\mathcal{M}$

- All non-empty models:

$$\mathcal{M} = \{M \subseteq \{1, \dots, p\}; M \neq \emptyset\},$$

  - only when $p \leq n$.

- All models containing the first variable:

$$\mathcal{M} = \{M \subseteq \{1, \dots, p\}; 1 \in M\},$$

  - only when $p \leq n$,
  - e.g. first variable is an intercept (first column of X composed of 1s).

- s-sparse models:

$$\mathcal{M} = \{M \subseteq \{1, \dots, p\}; |M| \leq s\},$$

  - allows for $n < p$,
  - $1 \leq s \leq n$ is the sparsity parameter.

Regression coefficients of interest

- The projection-based target: Let for $M \in \mathcal{M}$,

$$\beta_M^{(n)} = \underset{|M| \times 1 \text{ vector } v}{\operatorname{argmin}} \|\mu - X[M]v\|^2$$
$$= \left(X'[M]X[M]\right)^{-1} X'[M]\mu.$$

$\Longrightarrow$ Same as $\beta^\star$ above but for selected variables.

$\Longrightarrow$ $\beta_M^{(n)}$ is a target of inference in this talk.

$\Longrightarrow$ Motivated in [Berk et al., 2013].

$\Longrightarrow$ Subsequently considered in [Lee et al., 2016, Tibshirani et al., 2018],...

$\Longrightarrow$ When $p < n$ and $\mu \notin \mathrm{span}(X)$: links to extensive literature on misspecified parametric models [Eicker, 1967, Huber, 1967, White, 1982].

Illustration (1/2)

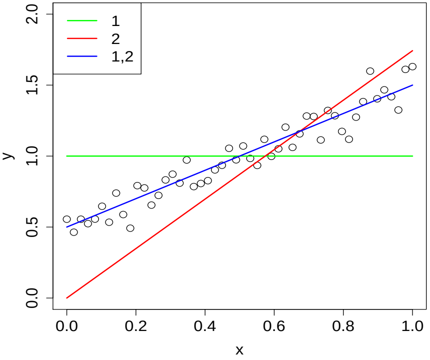
- $n = 50$, $p = 2$
- $$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$
- **well-specified case:**
$$\mu_i = 1/2 + x_i$$
for $i = 1, \ldots, n$.
  - $\mu \in \mathrm{span}(X)$.
- **misspecified case:**
$$\mu_i = -1/2 + x_i + 4x_i^2$$
for $i = 1, \ldots, n$.
  - $\mu \notin \mathrm{span}(X)$.

Illustration (2/2)

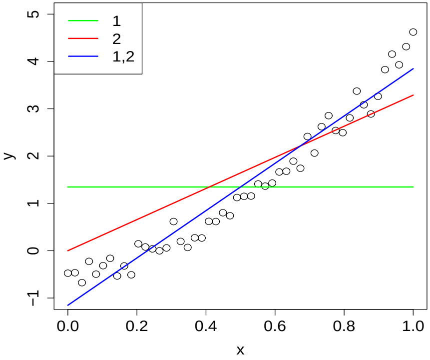
Plot of

- Observations $Y_1, \ldots, Y_n$,
- $\left(X[M]\beta_M^{(n)}\right)_i$, $i = 1, \ldots, n$, for

$$M = \{1\}, \quad M = \{2\} \text{ and } M = \{1, 2\}.$$



(well-specified)



(misspecified)

Post-model-selection inference

- Model selection procedure: data-driven selection of the model with

$$\hat{M}(Y) = \hat{M} \in \mathcal{M}.$$

  - Sequential testing, AIC, BIC, LASSO, SCAD [Fan and Li, 2001], MCP [Zhang, 2010],...

- In [Berk et al., 2013], target for inference is $\beta_{\hat{M}}^{(n)}$ and $\hat{M}$ can be any model selection procedure.
  - Model selector $\hat{M}$ is imposed.
  - Objective: best coefficients in this imposed model.

This is what we call the post-model-selection inference setting.

Discussion (1/2)

- Motivated by the following common practice in applications:
  1. select model $\hat{M}$ from data Y,
  2. apply usual confidence intervals/tests with design matrix $X[\hat{M}]$.

  $\Longrightarrow$ Invalid because $\hat{M}$ is data-dependent.

  $\Longrightarrow$ Aim at changing tests/confidence intervals so that they become valid.

- Motivation for considering target $\beta^{(n)}_{\hat{M}}$.
  - Best we can do once $\hat{M}$ is fixed.
  - Relevant in misspecified case when $\mu \notin \mathrm{span}(X)$ (when $p < n$).
  - Relevant when $p > n$ and no sparse representation of μ in X.

Discussion (2/2)

- Aim for procedures that work for any function $Y \mapsto \hat{M}(Y)$.
- In practice $\hat{M}$ can be
  - not formally defined,
  - imposed.
- Robustness to
  - hunting for significance,
  - also called p-hacking, data snooping,...
- This talk is not about how to select a “good” model $\hat{M}$.

Related literature

- We consider the setting of [Berk et al., 2013],
  - confidence intervals for $\beta^{(n)}_{\hat{M}}$ for any $\hat{M}$,
  - subsequent related work [Zhang, 2017, Kuchibhotla et al., 2020, Kuchibhotla et al., 2022], ...

- In [Lee et al., 2016, Tibshirani et al., 2018, Panigrahi and Taylor, 2022],...,
  - confidence intervals for $\beta^{(n)}_{\hat{M}}$,
  - $\hat{M}$ is specific: LASSO, sequential testing,...
  - valid (coverage probability) conditionally to $\hat{M}$.

- Hybridation of former 2 settings: [McCloskey, 2023].

- In [van de Geer et al., 2014], the LASSO model selector is used for confidence intervals in sparse well-specified models in high-dimension.

- Some intrinsic difficulties in post-model-selection inference were discussed earlier in [Leeb, 2005, Leeb and Pötscher, 2006], ...

Confidence intervals

We consider confidence intervals for

$$\left(\beta_{\hat{M}}^{(n)}\right)_j,$$

for $j = 1, \ldots, |\hat{M}|$, of the form

$$CI_{\hat{M},j} = \left(\hat{\beta}_{\hat{M}}\right)_j \pm K \|s_{\hat{M},j}\| \sigma,$$

with

$$s'_{\hat{M},j} = \text{row } j \text{ of } \left(X'[\hat{M}]X[\hat{M}]\right)^{-1} X'[\hat{M}].$$

Interpretation:

- For fixed M and j,

$$\left(\hat{\beta}_M\right)_j - \left(\beta_M^{(n)}\right)_j \sim \mathcal{N}(0, \|s_{M,j}\|^2 \sigma^2).$$

- Thus, selecting K as a Gaussian quantile is valid when M is deterministic.
- When $\hat{M}$ is random, K needs to be larger to account for model selection.

$\Longrightarrow$ Question: choosing K.

Reduction to a simultaneous coverage problem

[Berk et al., 2013].
The coverage

$$\text{for } j = 1, \ldots, |\hat{M}|, \quad \left(\beta^{(n)}_{\hat{M}}\right)_j \in CI_{\hat{M},j}$$

holds if the simultaneous coverage

$$\text{for } M \in \mathcal{M} \text{ and } j = 1, \ldots, |M|, \quad \left(\beta^{(n)}_{M}\right)_j \in CI_{M,j}$$

holds. This is equivalent to

$$\text{for } M \in \mathcal{M} \text{ and } j = 1, \ldots, |M|, \quad \frac{\left|\left(\hat{\beta}_M\right)_j - \left(\beta^{(n)}_M\right)_j\right|}{\|s_{M,j}\|\sigma} \leq K.$$

POSI (post-selection inference) constant

The last event can be rewritten as

$$\max_{\substack{M\in\mathcal{M},\\ j=1,\dots,|M|}} \left| \frac{s'_{M,j}}{\|s_{M,j}\|} \frac{(Y-\mu)}{\sigma} \right| \leq K.$$

- Distribution of the maximum does not depend on μ, σ.

$\Longrightarrow$ Taking

$$K = K_{1-\alpha}(X)$$

(POSI constant) as the $1-\alpha$ quantile of this maximum yields

$$\mathbb{P}\left(\text{for } j = 1, \dots, |\hat{M}|, \ \left(\beta^{(n)}_{\hat{M}}\right)_j \in CI_{\hat{M},j} \right) \geq 1-\alpha,$$

for all n, p, $\mu \in \mathbb{R}^n$, $\sigma > 0$.

$\Longrightarrow$ Uniformly valid confidence intervals [Berk et al., 2013].

$\Longrightarrow$ $K_{1-\alpha}(X)$ is optimal to guarantee this property.

POSI constant

$K_{1-\alpha}(X)$ quantile $1-\alpha$ of

$$\max_{\substack{M \in \mathcal{M}, \\ j=1,\ldots,|M|}} \left| \frac{s'_{M,j}}{\|s_{M,j}\|} (U/\sigma) \right|$$

with $U/\sigma \sim \mathcal{N}(0, I_n)$.

- Supremum norm of a large centered Gaussian vector,
  - dimension $\sum_{M \in \mathcal{M}} |M|$,
  - up to $p 2^{p-1}$ when $p \leq n$.
- With unit variances.
- Rank of covariance matrix $\leq \min(n, p)$.
- Alternatively: many one-dimensional projections of a standard Gaussian vector.

Computation of the POSI constant

- When p not too large, $K_{1-\alpha}(X)$ can be estimated by Monte Carlo,
  - say $p < 30$ when $\mathcal{M}$ is unrestricted,
  - larger p for sparse models,
  - `R` package `PoSI`.
- But cost usually exponential in p.
- Upper bound

$$B_{1-\alpha} \geq K_{1-\alpha}(X)$$

  suggested in [Berk et al., 2013], see also [Bachoc et al., 2018, Bachoc et al., 2020],
  - computation complexity $\approx$ constant w.r.t. n, p,
  - can be used in practice for large n, p.

How large are the POSI constant and its upper bound?

[Berk et al., 2013], see also [Bachoc et al., 2018, Bachoc et al., 2020].

- Fixed model, $\mathcal{M} = \{M_0\}$:

$$\sup_{X \ n \times p \ \text{matrix}} K_{1-\alpha}(X) = O(1).$$

- All models, $p \leq n$, $\mathcal{M} = \{M \subseteq \{1, \ldots, p\}\}$:

$$\inf_{X \ n \times p \ \text{matrix}} K_{1-\alpha}(X) = \sqrt{2 \log(p)}(1 + o(1)),$$

$$0.6363\sqrt{p}(1 + o(1)) \leq \sup_{X \ n \times p \ \text{matrix}} K_{1-\alpha}(X) \leq 0.866\sqrt{p}(1 + o(1)).$$

$\Longrightarrow K_{1-\alpha}(X)$ depends on X in a complex way.

Upper bound.

- Sparse models, $\mathcal{M} = \{M \subseteq \{1, \ldots, p\}; |M| \leq s\}$, $s \leq n$:

$$B_{1-\alpha} = O\left(\sqrt{s \log\left(\frac{p}{s}\right)}\right).$$

Linear predictors

This section is based on the paper:

Bachoc, F., Leeb, H., & Pötscher, B.M., Valid confidence intervals for post-model-selection predictors, *Annals of Statistics*, 47(3), 1475-1504, 2019.

- We consider a $p \times 1$ vector x_0,
  - new explanatory variables.
- Define $x_0[M]$: subvector of x_0 with indices in M,
  - for $M \in \mathcal{M}$.
- We want to cover the post-model-selection predictor

$$x_0[\hat{M}]'\beta_{\hat{M}}^{(n)}.$$

Confidence intervals

Adaptation of [Berk et al., 2013].

- Confidence interval

$$CI_{\hat{M},x_0} = x_0[\hat{M}]'\hat{\beta}_{\hat{M}} \pm K_{1-\alpha}(X,x_0)\|s_{\hat{M},x_0}\|\sigma,$$

- with

$$s'_{\hat{M},x_0} = x_0[\hat{M}]' \left(X'[\hat{M}]X[\hat{M}] \right)^{-1} X'[\hat{M}],$$

- with $K_{1-\alpha}(X,x_0)$ the $1-\alpha$ quantile of

$$\max_{M \in \mathcal{M}} \left| \frac{s'_{M,x_0}}{\|s_{M,x_0}\|} \frac{(Y-\mu)}{\sigma} \right|.$$

- We still have an upper bound

$$B'_{1-\alpha} \geq K_{1-\alpha}(X,x_0).$$

Case of partially observed x_0

- Can frequently happen that
  - x_0 not observed entirely,
  - only $x_0[\hat{M}]$ is observed,
  - variable selection for cost reasons.
- In this case $K_{1-\alpha}(X, x_0)$ is unavailable.
- We still have the upper bound $B'_{1-\alpha}$.
- We also suggest

$$K_{2,1-\alpha}(X, x_0[\hat{M}], \hat{M}) = \sup_{x_0[\hat{M}^c]} K_{1-\alpha}(X, x_0),$$

  - very hard to compute,
  - but theoretically interesting.

$\Longrightarrow$

$$K_{1-\alpha}(X, x_0) \leq K_{2,1-\alpha}(X, x_0[\hat{M}], \hat{M}) \leq B'_{1-\alpha}.$$

Large p analysis for orthogonal design matrices (1/2)

- When X has orthogonal columns, $K_{1-\alpha}(X)$ has rate $\sqrt{\log(p)}$ [Berk et al., 2013].
- From that we deduce that $K_{1-\alpha}(X, x_0)$ has rate $\sqrt{\log(p)}$ when x_0 is a sequence of basis vectors.

Large p analysis for orthogonal design matrices (2/2)

Proposition

Let $\mathcal{M}$ be the power set of $\{1, ..., p\}$ (minus empty set).
(a) Let X have orthogonal columns. There exists a sequence of vectors x_0 such that $K_{1-\alpha}(X, x_0)$ satisfies

$$\liminf_{p\to\infty} K_{1-\alpha}(X, x_0)/\sqrt{p} \geq 0.63.$$

(b) Let $\gamma \in [0, 1)$ be given. Then $K_{2,1-\alpha}(X, x_0[M], M)$ satisfies

$$\liminf_{p\to\infty} \inf_{x_0\in\mathbb{R}^p} \inf_{X\in\mathrm{X}(p)} \inf_{M\in\mathcal{M},|M|\leq\gamma p} K_{2,1-\alpha}(X, x_0[M], M)/\sqrt{p} \geq 0.63\sqrt{1-\gamma},$$

where $\mathrm{X}(p) = \bigcup_{n\geq p} \{X : X \text{ is } n \times p \text{ with non-zero orthogonal columns}\}$.

$\Longrightarrow$ Strong impact of x_0 on $K_{1-\alpha}(X, x_0)$.
$\Longrightarrow$ Price to pay when only $x_0[\hat{M}]$ is observed.

Summary of another contribution of the paper

- We consider the random regressors setting.
- The rows of X and x_0 are realizations from a distribution $\mathcal{L}$.
- We define the post-model-selection predictor

$$x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$$

defined based on $\mathcal{L}$ rather than on X.

- We show that the same confidence intervals as before work asymptotically.
    - p fixed, $n \to \infty$ here.

$\Longrightarrow$ Recent work on random regressors, [Buja et al., 2019, Kuchibhotla et al., 2021].

The paper

This section is based on the paper:

Bachoc, F., Preinerstorfer, D. & Steinberger, L., Uniformly valid confidence intervals post-model-selection, *Annals of Statistics*, 48(1), 440-463, 2020.

Data and models

Data.

- We consider a triangular array of independent $1 \times l$ random vectors $y_{1,n}, ..., y_{n,n}$.
- We let $\mathbb{P}_n = \bigotimes_{i=1}^n \mathbb{P}_{i,n}$ be the distribution of $y_n = (y'_{1,n}, \ldots, y'_{n,n})'$, where $\mathbb{P}_{i,n}$ is the distribution of $y_{i,n}$.

Models.

- We now consider a set $\mathsf{M}_n = \{\mathbb{M}_{1,n}, \ldots, \mathbb{M}_{d,n}\}$ composed of d models.
- $\mathbb{M}_{i,n}$ is a set of distributions on $\mathbb{R}^{n \times l}$.
- d does not depend on n (fixed-dimensional asymptotics).

$\Longrightarrow$ We do not assume that the observation distribution $\mathbb{P}_n$ belongs to one of the $\{\mathbb{M}_{1,n}, \ldots, \mathbb{M}_{d,n}\}$. The set of models can be misspecified.

Parameters and estimators

Parameters.

- We define for each model $\mathbb{M} \in \mathsf{M}_n$ an optimal parameter $\theta^*_{\mathbb{M},n} = \theta^*_{\mathbb{M},n}(\mathbb{P}_n)$, that we assume to be non-random and of fixed dimension $m(\mathbb{M})$.
- In the case of linear models:
  - each $\mathbb{M} \in \mathsf{M}_n$ corresponds to a $M \subseteq \{1, \dots, p\}$,
  - $\theta^*_{\mathbb{M},n} = \beta^{(n)}_M$.
- The optimal parameter $\theta^*_{\mathbb{M},n}$ is specific to the model $\mathbb{M}$.

Estimators.

- We consider, for each $\mathbb{M} \in \mathsf{M}_n$, an estimator $\hat{\theta}_{\mathbb{M},n}$ of the optimal parameter $\theta^*_{\mathbb{M},n}$.

Post-model-selection inference

Model selection.

- We consider a model selection procedure: a function $\hat{\mathbb{M}}_n : \mathbb{R}^{n \times l} \to \mathsf{M}_n$.
- We are hence interested in constructing confidence intervals for the random quantity of interest $\theta^*_{\hat{\mathbb{M}}_n, n}$.

Main idea and notation

Main idea.

- We aim at showing a joint asymptotic normality of $\{\hat{\theta}_{\mathbb{M},n} - \theta^*_{\mathbb{M},n}\}_{\mathbb{M} \in \mathsf{M}_n}$.
- We then use the same construction as in the Gaussian linear case for the confidence intervals.
- Additional difficulty: we do not know the asymptotic covariance matrix.

Notation.

- $\hat{\theta}_n = (\hat{\theta}'_{\mathbb{M}_1,n}, \ldots, \hat{\theta}'_{\mathbb{M}_d,n})'$.
- $\theta^*_n = (\theta^{*'}_{\mathbb{M}_1,n}, \ldots, \theta^{*'}_{\mathbb{M}_d,n})'$.
- Let $k = \sum_{j=1}^{d} m(\mathbb{M}_{j,n})$ be the dimension of $\hat{\theta}_n$.

Joint asymptotic normality

Assumption: linear approximation

$$\hat{\theta}_n - \theta_n^* = \overbrace{\sum_{i=1}^{n} \underbrace{g_{i,n}(y_{i,n})}_{\text{centered}}}^{:=r_n} \; + \; \text{negligible.}$$

- Let d_w be a distance generating the topology of weak convergence for distributions on an Euclidean space.
- Let $\mathrm{corr}(\Sigma)$ be the correlation matrix obtained from a covariance matrix Σ.
- Let $\mathrm{diag}(\Sigma)$ be obtained by setting the off-diagonal elements of Σ to 0.

Lemma

$$d_w \left(\textit{law of } \mathrm{diag}(\mathbb{V}\mathbb{C}_n(r_n))^{-1/2} \left(\hat{\theta}_n - \theta_n^* \right), \mathcal{N}(0, \mathrm{corr}(\mathbb{V}\mathbb{C}_n(r_n))) \right) \to 0.$$

Some notation

- For $\alpha \in (0,1)$ and for a covariance matrix Γ, let $K_{1-\alpha}(\Gamma)$ be the $1-\alpha$-quantile of $\|Z\|_\infty$ for $Z \sim N(0,\Gamma)$.
  $\Longrightarrow$ Very similar to above POSI constant.

- For $\mathbb{M} = \mathbb{M}_{q,n} \in \mathsf{M}_n$ let

$$\rho(\mathbb{M}) \quad := \quad \sum_{\ell=1}^{q-1} m(\mathbb{M}_{\ell,n}).$$

$\Longrightarrow$ $\rho(\mathbb{M}) + j$ is the index of $(\theta^{*'}_{\mathbb{M},n})_j$ in $(\theta^{*'}_{\mathbb{M}_1,n}, \ldots, \theta^{*'}_{\mathbb{M}_d,n})'$
for $j \in \{1, \ldots, m(\mathbb{M})\}$.

Let $\alpha \in (0,1)$. Let $\hat{S}_n$ be such that, with $\|A\|$ the largest singular value of A,

$$\| \operatorname{corr}(\hat{S}_n) - \operatorname{corr}\left(\mathbb{VC}_n(r_n)\right) \| + \| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(\hat{S}_n) - I_k \| \to_p 0.$$

Consider, for $\mathbb{M} \in \mathsf{M}_n$ and $j = 1, \dots, m(\mathbb{M})$, the confidence interval

$$\mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{est}} \quad = \quad \left[\hat{\theta}_{\mathbb{M},n}\right]_j \pm \sqrt{[\hat{S}_n]_{\rho(\mathbb{M})+j,\rho(\mathbb{M})+j}}\; K_{1-\alpha}\left(\operatorname{corr}(\hat{S}_n)\right).$$

Theorem

Then, $\mathbb{P}_n\left(\left[\theta^*_{\mathbb{M},n}\right]_j \in \mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{est}} \text{ for all } \mathbb{M} \in \mathsf{M}_n \text{ and } j = 1, \dots, m(\mathbb{M})\right)$ *goes to* $1-\alpha$ *as* $n \to \infty$*. In particular, for any model selection procedure* $\hat{\mathbb{M}}_n$*, we have*

$$\liminf_{n\to\infty} \mathbb{P}_n\left(\left[\theta^*_{\hat{\mathbb{M}}_n,n}\right]_j \in \mathrm{CI}_{1-\alpha,\hat{\mathbb{M}}_n}^{(j),\mathrm{est}} \text{ for all } j = 1, \dots, m(\hat{\mathbb{M}}_n)\right) \geq 1-\alpha.$$

Confidence intervals with conservative estimator of asymptotic covariance matrix

- When the models are misspecified it may not be possible to estimate $\mathbb{VC}_n(r_n)$ consistently.

Over-estimation of the diagonal components of $\mathbb{VC}_n(r_n)$

1. Recall the linear approximation

$$\hat{\theta}_n - \theta_n^* = \sum_{i=1}^{n} \underbrace{g_{i,n}(y_{i,n})}_{\substack{\text{centered} \\ \text{not observed}}} \; + \; \text{negligible.}$$

2. Consider computable $\tilde{g}_{i,n}(y_n)$ such that

$$\tilde{g}_{i,n}(y_n) = g_{i,n}(y_{i,n}) \; + \; \text{deterministic bias} \; + \; \text{negligible.}$$

3. Take empirical second moments of $(\tilde{g}_{i,n}(y_n))_{i=1,\ldots,n}$.

- Also there exists an upper-bound of $K_{1-\alpha}\left(\mathrm{corr}(\mathbb{VC}_n(r_n))\right)$ (similar to $B_{1-\alpha}$).

$\Longrightarrow$ We obtain similar asymptotic guarantees as before with more conservative confidence intervals.

Applications

- We have seen a general method that can be applied to specific situations on a case by case basis.
- Need uniform central limit theorems for fixed models in misspecified cases (sandwich rule).
- Need to consistently overestimate variances.
- In the paper, we provide applications to
  - homoscedastic linear models with homoscedastic data,
  - heteroscedastic linear models with heteroscedastic data,
  - binary regression models with binary data.

Binary regression: data

Data.

- $l = 1$: scalar observations.
- $n \times 1$ observation vector

$$y_n = \begin{pmatrix} y_{1,n} \\ \vdots \\ y_{n,n} \end{pmatrix}.$$

  - Independent components.
  - $y_{i,n} \in \{0, 1\}$.
  - For $i = 1, ..., n$, $\mathbb{P}(y_{i,n} = 1) \in [\delta, 1 - \delta]$ for fixed $\delta > 0$ (technical for asymptotics).

$\Longrightarrow \mathbb{P}_n$ is a distribution on $\{0, 1\}^n$ with independent components and non-vanishing 'randomness'.

Binary regression: generalized linear models

Models.

- Let X be a $n \times p$ design matrix.
- Let X_i be the ith row of X.
- Model $\mathbb{M}$ identified by set of variables $M \in \mathcal{M} \subseteq \{M \subseteq \{1, \dots, p\}\}$.
- Under model $\mathbb{M}$ we assume that for $i = 1, ..., n$

$$\mathbb{P}(y_{i,n} = 1) = \frac{e^{X_i[M]\theta_{\mathbb{M}}}}{1 + e^{X_i[M]\theta_{\mathbb{M}}}}. \tag{1}$$

  - Canonical link function.
  - For some $|M| \times 1$ vector $\theta_{\mathbb{M}}$.
  - With $X_i[M]$ the ith row of $X[M]$.

$\Longrightarrow$ $\mathbb{M}$ is the set of distributions on $\mathbb{R}^n$ with independent components in $\{0, 1\}$ and with mean vector given by (1).

Binary regression: target and estimator

Target.

- For a model $\mathbb{M}$

$$\theta^*_{\mathbb{M},n} \in \underset{\theta_{\mathbb{M}} \in \mathbb{R}^{|M|}}{\operatorname{argmin}} \mathrm{KL}(\mathbb{P}_n \parallel \mathbb{P}_{\mathbb{M},\theta_{\mathbb{M}}}),$$

  with

  - $\mathbb{P}_{\mathbb{M},\theta_{\mathbb{M}}}$ the distribution in model $\mathbb{M}$ with parameter $\theta_{\mathbb{M}}$,
  - $\mathbb{P}_n$ the true distribution of the observation vector.

Estimator.

- $\hat{\theta}_{\mathbb{M},n}$: the maximum likelihood estimator in the model $\mathbb{M}$.

$\Longrightarrow$ We show unicity of the target and uniform consistency and unicity (with probability $\to 1$) of the estimator.

$\Longrightarrow$ Related work [Fahrmeir, 1990, Lv and Liu, 2014].

Binary regression: over-estimation of covariance matrix

- Linearization:

$$\hat{\theta}_{\mathbb{M},n} - \theta^*_{\mathbb{M},n} = \overbrace{\left[\mathbb{E}_n(H^*_{\mathbb{M},n})\right]^{-1} \sum_{i=1}^{n} X_i[\mathbb{M}]' \left(y_{i,n} - \mathbb{E}_n(y_{i,n})\right)}^{r_n} + \text{negligible}$$

with $H^*_{\mathbb{M},n}$ the Hessian of $-\log(\text{likelihood})$ for model $\mathbb{M}$ at $\theta^*_{\mathbb{M},n}$.

- Over-estimator of diagonal block of $\mathbb{VC}_n(r_n)$ corresponding to $\mathbb{M}$:

$$\left[\widehat{H}_{\mathbb{M},n}\right]^{-1} \left(\sum_{i=1}^{n} X_i[\mathbb{M}]' X_i[\mathbb{M}] \left(y_{i,n} - \widehat{y}_{\hat{\theta}_{\mathbb{M},n},i,n} \right)^2 \right) \left[\widehat{H}_{\mathbb{M},n}\right]^{-1}$$

with

- $\widehat{H}_{\mathbb{M},n}$ the Hessian at $\hat{\theta}_{\mathbb{M},n}$,
- $\widehat{y}_{\hat{\theta}_{\mathbb{M},n},i,n} = e^{X_i[\mathbb{M}]\hat{\theta}_{\mathbb{M},n}} / (1 + e^{X_i[\mathbb{M}]\hat{\theta}_{\mathbb{M},n}})$.

Some simulation results

In a Monte Carlo simulation (1000 repetitions) for logistic regression ($p = 10, n = 30, 100$), we compare

- CI coverage for a nominal level at 0.9 (cov. 0.9),
- CI median length (med.),
- CI 90% quantile length (qua.)

for

- our post-selection inference CI (P),
- the CI by [Taylor and Tibshirani, 2017], specific to the lasso (L),
- the naive CI that ignores the presence of model selection (N).

model	cov. 0.9			med.			qua.		
selector	P	L	N	P	L	N	P	L	N
lasso (1)	0.99	0.89	0.84	4.26	7.44	2.09	6.97	43.33	3.42
lasso (2)	1.00	0.85	0.68	1.63	2.31	0.74	1.90	13.52	0.84
lasso (3)	1.00	0.25	0.98	2.22	1.23	1.01	2.83	3.50	1.24
sig. hun.	0.95		0.39	4.40		2.63	6.22		3.63

Some simulation results in high dimension (1/2)

Monte Carlo simulation (1000 repetitions) for homoscedastic linear models ($p = 1000, n = 50$).

- The model selector is forward stepwise.

We compare

- CI coverage for a nominal level at 0.9 (cov.),
- CI median length (med.),
- CI 90% quantile length (qua.)

for

- our post-selection inference CI (P),
- the CI by [Tibshirani et al., 2016], specific to forward-stepwise (FS),
- the naive CI that ignores the presence of model selection (N).

Some simulation results in high dimension (2/2)

	Step 1			Step 2			Step 3			Simult.
	cov.	med.	qua.	cov.	med.	qua.	cov.	med.	qua.	cov.
P	0.99	8.33	9.38	1.00	10.39	12.73	1.00	11.49	14.35	0.99
FS	0.94	11.66	55.76	0.88	786.92	Inf	0.90	1754.00	Inf	0.77
N	0.58	3.54	3.98	0.49	3.33	4.08	0.45	3.22	4.03	0.08
P	0.91	7.24	8.07	1.00	9.34	12.15	1.00	10.36	13.68	0.91
FS	0.93	15.15	72.67	0.88	752.74	Inf	0.90	1582.32	Inf	0.76
N	0.00	3.07	3.43	0.12	3.00	3.90	0.19	2.91	3.84	0.00

Remarks.

- Top 3 rows: design matrix X has independent columns.
- Bottom 3 rows: design matrix X has correlated columns.
- The CI's P and FS use the knowledge that k variables are selected at step k.

Conclusion and perspectives

Conclusion.

- Inference for targets that depend on selected models.
- Simultaneous coverage of many correlated and normalized errors.
- Exact for Gaussian case $\Longrightarrow$ asymptotic for more general cases.
- R code of all experiments on personal GitHub page.

Personal subsequent work.

- Post-clustering inference, [Bachoc et al., 2023] see also [Gao et al., 2022].
- Inference post-selection of regions (ongoing), see also [Benjamini et al., 2019, Chernozhuokov et al., 2022].

Thank you for your attention!

Bibliography I

Bachoc, F., Blanchard, G., and Neuvial, P. (2018).
On the post selection inference constant under restricted isometry properties.
Electronic Journal of Statistics, 12(2).

Bachoc, F., Maugis-Rabusseau, C., and Neuvial, P. (2023).
Selective inference after convex clustering with l1 penalization.
arXiv:2309.01492.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2020).
Uniformly valid confidence intervals post-model-selection.
Annals of Statistics, 48(1):440–463.

Benjamini, Y., Taylor, J., and Irizarry, R. A. (2019).
Selection-corrected statistical inference for region detection with high-throughput assays.
Journal of the American Statistical Association, 114(527):1351–1365.

Bibliography II

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013).
Valid post-selection inference.
Annals of Statistics, 41(2):802–837.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., and Zhao, L. (2019).
Models as approximations ii: A model-free theory of parametric regression.
Statistical Science, 34(4):545–565.

Chernozhuokov, V., Chetverikov, D., Kato, K., and Koike, Y. (2022).
Improved central limit theorem and bootstrap approximations in high dimensions.
Annals of Statistics, 50(5):2562–2586.

Bibliography III

Eicker, F. (1967).
Limit theorems for regressions with unequal and dependent errors.
In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82. Berkeley, CA: University of California Press.

Fahrmeir, L. (1990).
Maximum likelihood estimation in misspecified generalized linear models.
Statistics, 21(4):487–502.

Fan, J. and Li, R. (2001).
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American statistical Association, 96(456):1348–1360.

Gao, L. L., Bien, J., and Witten, D. (2022).
Selective inference for hierarchical clustering.
Journal of the American Statistical Association, pages 1–11.

Bibliography IV

Huber, P. (1967).
The behavior of maximum likelihood estimates under nonstandard conditions.
In *Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967*.

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I., and Zhao, L. H. (2020).
Valid post-selection inference in model-free linear regression.
Annals of Statistics, 48(5):2953–2981.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2021).
Uniform-in-submodel bounds for linear regression in a model-free framework.
Econometric Theory, page 1–47.

Bibliography V

Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022).
Post-selection inference.
Annual Review of Statistics and Its Application, 9:505–527.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016).
Exact post-selection inference, with application to the lasso.
Annals of Statistics, 44(3):907–927.

Leeb, H. (2005).
The distribution of a linear predictor after model selection: conditional finite-sample distributions and asymptotic approximations.
Journal of Statistical Planning and Inference, 134:64–89.

Leeb, H. and Pötscher, B. M. (2006).
Can one estimate the conditional distribution of post-model-selection estimators?
Annals of Statistics, 34(5):2554 – 2591.

Bibliography VI

Lv, J. and Liu, J. S. (2014).
Model selection principles in misspecified models.
Journal of the Royal Statistical Society Series B, 76:141–167.

McCloskey, A. (2023).
Hybrid confidence intervals for informative uniform asymptotic inference after model selection.
Biometrika.

Panigrahi, S. and Taylor, J. (2022).
Approximate selective inference via maximum likelihood.
Journal of the American Statistical Association, pages 1–11.

Taylor, J. and Tibshirani, R. (2017).
Post-selection inference for l1-penalized likelihood models.
Canadian Journal of Statistics, pages 1–21.

Bibliography VII

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018).
Uniform asymptotic inference and the bootstrap after model selection.
Annals of Statistics, 46(3):1255 – 1287.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016).
Exact post-selection inference for sequential regression procedures.
Journal of the American Statistical Association, 111(514):600–620.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014).
On asymptotically optimal confidence regions and tests for high-dimensional models.
Annals of Statistics, **42**:1166–1202.

White, H. (1982).
Maximum likelihood estimation of misspecified models.
Econometrica, pages 1–25.

Bibliography VIII

Zhang, C. H. (2010).
Nearly unbiased variable selection under minimax concave penalty.
Annals of Statistics, 38(2):894–942.

Zhang, K. (2017).
Spherical cap packing asymptotics and rank-extreme detection.
IEEE Transactions on Information Theory, 63(7):4572–4584.